

Learning to Disentangle Dispersion and Disparity: A Dual-Pixel CASSI for Joint Spectral-Depth Imaging

Xiaodong Wang^{1,2}, Siming Zheng³, and Xin Yuan²

¹ Zhejiang University, Hangzhou, China

² Westlake University, Hangzhou, China

{wangxiaodong, xyuan}@westlake.edu.cn

³ vivo Mobile Communication Co., Ltd, Hangzhou, China
zhengsiming@vivo.com

Abstract. Snapshot compressive imaging, such as Coded Aperture Snapshot Spectral Imaging (CASSI), has achieved remarkable success in capturing dynamic hyperspectral scenes. However, most existing CASSI systems inherently assume a flat 2D scene, ignoring the 3D geometric structure. In real-world scenarios, depth variations cause spatially-varying defocus blur, which severely tangles with spectral dispersion and degrades reconstruction fidelity. Furthermore, capturing both hyperspectral and depth information typically requires bulky multi-camera setups. In this paper, we propose DP-CASSI, a novel computational imaging system that seamlessly integrates a Dual-Pixel (DP) sensor into the CASSI optical path. By capturing two phase-shifted half-aperture views in a single snapshot, DP-CASSI uniquely encodes both spectral dispersion and depth-dependent disparity. To recover these measurements, we formulate the forward physical model of DP-CASSI and introduce a unified end-to-end reconstruction network, *SpecNet*, which directly outputs both the hyperspectral cube and the depth map from the dual-pixel coded measurements. Extensive experiments on simulated datasets demonstrate that our method significantly outperforms existing single-task approaches, achieving state-of-the-art fidelity in both hyperspectral image reconstruction and depth estimation from a single snapshot.

Keywords: Hyperspectral reconstruction · Depth estimation

1 Introduction

Hyperspectral imaging (HSI) provides dense spectral measurements at every spatial location and enables applications in computational photography, material analysis, remote sensing, and scientific imaging. Yet the acquisition of a full hyperspectral datacube remains challenging because the signal is inherently high-dimensional, and conventional scanning systems are too slow for dynamic scenes. Snapshot compressive imaging [28] addresses this limitation by optically multiplexing the spectral volume into a single low-dimensional measurement that is later computationally reconstructed. Among snapshot systems, Coded Aperture Snapshot Spectral Imaging (CASSI) is one of the most influential designs because it combines coded modulation and wavelength-dependent dispersion in a compact optical pipeline, enabling single-shot HSI capture [28].

Despite this success, most existing CASSI systems and reconstruction methods are still derived under a simplified flat-scene assumption. Real scenes are three-dimensional, however, and points at different depths are projected with different defocus characteristics. In a CASSI system, this depth-dependent blur is not an independent nuisance: it becomes entangled with coded modulation and spectral dispersion, making the inverse problem considerably more ambiguous than in the idealized 2D case. Consequently, methods trained or designed under the flat-scene assumption degrade in scenes with significant depth variation.

This observation reveals a broader opportunity. Scene geometry should not be treated only as corruption; it can also serve as a useful cue for disentangling the measurement. Existing work on spectral compressive imaging usually reconstructs only the HSI cube, while geometry is either ignored or supplied by an external sensor. Conversely, systems that jointly recover spectrum and geometry often rely on multiple cameras, stereo hardware, or diffractive optical elements [21, 25, 8, 19]. These designs increase calibration complexity and reduce the compactness that makes snapshot imaging attractive.

In this paper, we propose *DP-CASSI* (Fig. 1), a snapshot computational imaging system that augments the CASSI pipeline with a Dual-Pixel (DP) sensor. A DP sensor splits each pixel into two half-aperture sub-pixels and records two phase-shifted views in a single exposure. This mechanism has proven useful for autofocus, depth reasoning, and defocus analysis in standard photography [10, 18, 22, 1]. Our key idea is to embed this capability directly into the CASSI optical path. The resulting system captures two coded and spectrally dispersed sub-aperture measurements that jointly encode spectral dispersion and depth-dependent disparity. Unlike prior works that reconstruct spectrum and depth separately, we treat the problem as a unified inverse mapping and solve it with a single end-to-end network, *SpecNet*, which directly outputs both the hyperspectral cube and the depth map.

Our approach makes three main technical contributions. First, we formulate a dual-pixel CASSI forward model that explicitly accounts for coded aperture modulation, spectral shearing, half-aperture imaging, and depth-dependent defocus. Second, we build a fully differentiable simulation pipeline that starts from an HSI cube and corresponding depth map, generates realistic left/right DP views following the physically motivated model of Abuolaim et al. [1], and then applies the CASSI shift-and-sum measurement process to produce the final coded measurements. Third, we design a straightforward end-to-end reconstruction network, *SpecNet*, which shares features across scales but does not rely on explicit cross-task interactions; it jointly predicts hyperspectral bands and depth from the DP-CASSI measurements. Experiments on simulated HSI-depth datasets show that DP-CASSI improves spectral reconstruction and enables reliable depth estimation from a single snapshot.

2 Related Work

2.1 CASSI Reconstruction Algorithms

CASSI reconstruction is a severely ill-posed inverse problem. Early methods mainly relied on handcrafted priors and iterative optimization, including TV-based reconstruction,

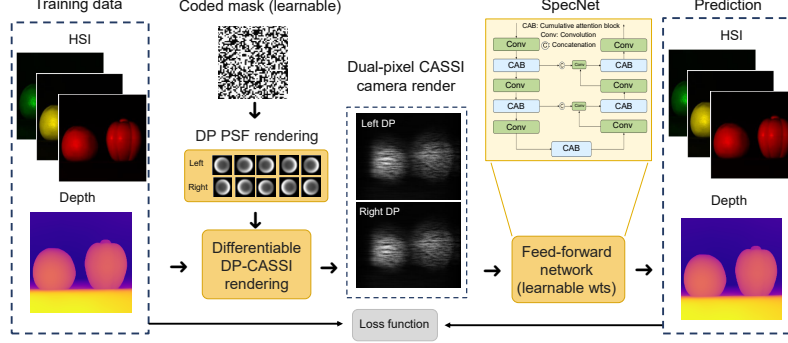


Fig. 1. DP-CASSI pipeline for joint hyperspectral and depth recovery from coded dual-pixel measurements.

spectral-spatial regularization, and low-rank or nonlocal modeling [27, 9, 14]. These approaches are physically interpretable but often produce over-smoothed results and require long runtimes. Learning-based methods significantly improved reconstruction quality and efficiency. λ -Net [17] demonstrated that compact CNNs can directly learn the mapping from coded measurements to HSI cubes. TSA-Net [16] introduced spatial-spectral self-attention, while later model-driven approaches such as DAUHST [6] combined unfolding with stronger learned priors. Transformer-based methods have become especially prominent: MST [4], MST++ [5], CST [3], S2Former [20] and SPECAT [26] improve long-range spectral modeling and efficiently exploit mask-aware structure. However, these methods typically assume 2D image formation, overlooking the coupling between spectral dispersion and depth-dependent blur.

2.2 Dual-Pixel Imaging and Depth Estimation

Dual-pixel sensing has been studied extensively for geometric inference and computational photography. Garg et al. [10] showed that DP images can support single-camera depth estimation by exploiting phase-shifted sub-aperture views. Pan et al. [18] further modeled DP formation for joint depth estimation and image restoration, revealing that defocus and disparity can be mutually informative. Xin et al. [22] jointly estimated defocus maps and deblurred images from a single DP image, while Kurita et al. [13] revisited DP disparity estimation from a lightweight physics-informed perspective. These works confirm that DP sensors encode meaningful geometry in a single shot. Abuolaim et al. [1] further modeled asymmetric half-aperture point-spread functions to synthesize realistic DP observations from image and depth data.

2.3 Joint Spectral and Depth Reconstruction

Joint recovery of spectral and geometric information has attracted increasing attention because these modalities provide complementary cues. Early systems relied on multiple sensors or views. Wang et al. [21] used a cross-modal stereo setup to jointly estimate depth and hyperspectral data, and Yao et al. [25] combined a low-resolution spectral camera with a high-resolution RGB camera to recover spectrum and depth. More compact snapshot solutions have also emerged. Baek et al. [2] proposed a single-shot hyperspectral-depth system using learned diffractive optics, and Ding et al. [8] introduced a snapshot depth-spectral architecture based on image mapping and light-field principles. Related systems include hyperspectral light-field imaging [23, 7] and specialized hyperspectral-depth imagers for surface measurement [15].

Our setting differs from prior work in two key aspects. First, we focus on the CASSI paradigm, where depth is not auxiliary information but is intrinsically entangled with coded spectral dispersion, which further degraded the imaging process due to the defocus effect. Second, DP-CASSI enables single-shot, plug-and-play spectral-depth imaging without multiple cameras, repeated captures, or complex diffractive designs.

3 Method

3.1 DP-CASSI Image Formation

Figure 2 summarizes the optical layout and signal flow of DP-CASSI. Let $\mathbf{X} \in \mathbf{R}^{H \times W \times L}$ denote the latent hyperspectral image (HSI) cube, where H , W , and L are the spatial height, spatial width, and number of spectral bands, respectively. Let $\mathbf{D} \in \mathbf{R}^{H \times W}$ denote the corresponding depth map. For the λ -th wavelength band, we write the spectral slice as $\mathbf{X}_\lambda \in \mathbf{R}^{H \times W}$.

In a conventional coded aperture snapshot spectral imaging (CASSI) system, a hyperspectral image (HSI) cube $\mathbf{X} \in \mathbf{R}^{H \times W \times L}$, with L spectral bands, is modulated by a coded aperture $\mathbf{M} \in \mathbf{R}^{H \times W}$ and then dispersed by a wavelength-dependent prism before integration on a 2D detector. The measurement at the sensor can be written as

$$\mathbf{Y} = \sum_{\lambda=1}^L \mathcal{S}_\lambda(\mathbf{M} \odot \mathbf{X}_\lambda) + \mathbf{N}, \quad (1)$$

where $\mathbf{X}_\lambda$ is the λ -th spectral slice, $\mathcal{S}_\lambda(\cdot)$ is the prism-induced horizontal shift, $\odot$ denotes element-wise multiplication, and $\mathbf{N}$ models noise. In discrete pixel form, this reads

$$\mathbf{Y}(i, j) = \sum_{\lambda=1}^L \mathbf{M}(i, j - \Delta_\lambda) \mathbf{X}_\lambda(i, j - \Delta_\lambda) + \mathbf{N}(i, j), \quad (2)$$

where Δ_λ is the pixel displacement for band λ due to spectral dispersion. This formulation highlights that the detector collapses the 3D HSI cube into a 2D coded measurement, combining spectral information from all wavelengths.

CASSI allows snapshot acquisition of hyperspectral scenes, but the resulting measurement is underdetermined, requiring specialized reconstruction algorithms to recover the HSI cube from $\mathbf{Y}$.

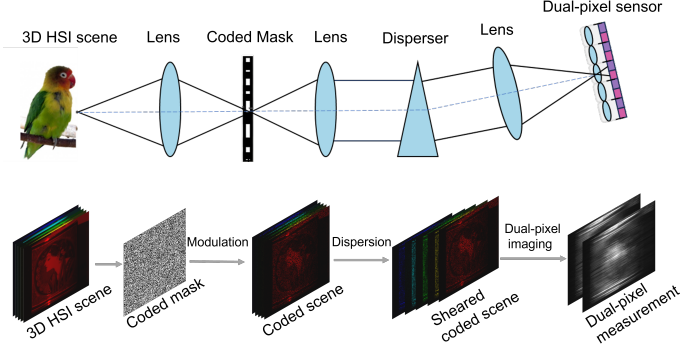


Fig. 2. Optical and signal-flow schematic of DP-CASSI, where coded modulation, spectral dispersion, and dual-pixel sensing jointly encode spectrum and depth.

3.2 Dual-pixel PSF Rendering

Figure 3 illustrates the dual-pixel PSF formation and disparity. To synthesize realistic dual-pixel (DP) observations, we follow a thin-lens approximation together with the parametric DP-PSF model in prior work [1]. Let f denote the focal length, d the aperture diameter, u the distance from a scene point to the lens, and v the corresponding image distance. For a point lying on the focal plane, its object distance is denoted by s . Under the thin-lens model, the focused image distance is

$$v_s = \frac{fs}{s - f}, \quad (3)$$

while a point at depth u forms an image at

$$v = \frac{fu}{u - f}. \quad (4)$$

The radius of the circle of confusion (CoC) is then approximated by

$$r = \frac{d}{2} \left| \frac{v - v_s}{v} \right|. \quad (5)$$

Given the CoC radius r , we adopt the Butterworth-based parametric DP-PSF model. Specifically, a 2D Butterworth profile is defined as

$$B(x, y) = \left(1 + \left(\frac{D_o}{\sqrt{(x - x_0)^2 + (y - y_0)^2}} \right)^{2n} \right)^{-1}, \quad (6)$$

where n is the filter order and D_o controls the cutoff. The combined DP kernel is modeled as

$$H = B \odot C(x_0, y_0), \quad (7)$$

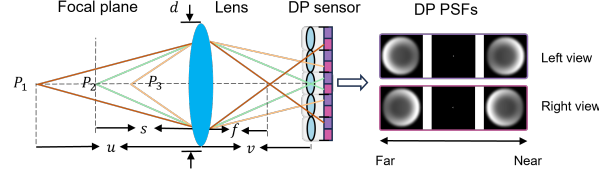


Fig. 3. Dual-pixel PSF formation and disparity for in-focus and defocused points.

where $C(x_0, y_0)$ is a circular disk kernel with radius r , and $\circ$ denotes element-wise multiplication. The combined kernel is decomposed into left- and right-view sub-aperture PSFs; points in front of and behind the focal plane produce opposite disparity directions in the two DP views while remaining fully differentiable.

Given the combined DP kernel H , we further split it into left and right sub-aperture kernels, denoted by H^L and H^R , respectively. For each spectral slice $\mathbf{X}_\lambda$, the corresponding dual-pixel observations are obtained by spatial convolution:

$$\mathbf{X}_\lambda^s = H^s * \mathbf{X}_\lambda, \quad s \in \{L, R\}, \quad (8)$$

where $*$ denotes 2D convolution. This indicates a spatially-varying spectral convolution operation.

3.3 Differentiable Data Generation

After dual-pixel rendering, each spectral slice is decomposed into left and right sub-aperture observations, denoted by $\mathbf{X}_\lambda^L$ and $\mathbf{X}_\lambda^R$. These two views are then modulated by the coded aperture and shifted by the dispersive element in the same way as in conventional CASSI. The resulting dual-pixel coded measurements are written as

$$\mathbf{Y}^s = \sum_{\lambda=1}^L \mathcal{S}_\lambda(\mathbf{M} \odot \mathbf{X}_\lambda^s) + \mathbf{N}^s, \quad s \in \{L, R\}, \quad (9)$$

where $\mathbf{X}_\lambda^s$ denotes the DP-rendered spectral slice for the left/right branch, $\mathbf{M}$ is the coded aperture, $\mathcal{S}_\lambda(\cdot)$ is the wavelength-dependent dispersion operator, and $\mathbf{N}^s$ models sensor noise. Because paired DP-CASSI measurements, HSI cubes, and dense depth maps are difficult to capture in the real world, we train the model using the differentiable synthetic pipeline in Fig. 4. We convert each HSI cube into a realistic RGB observation, estimate dense pseudo-depth using DepthAnything-V2 [24], render wavelength-wise left/right DP observations, and apply the CASSI forward model in Eq. 9 to generate coded measurements.

DepthAnything-V2 is used as relative pseudo-depth rather than metric depth. Depth maps serve two roles in our pipeline: they drive dual-pixel blur simulation, where normalized depth in $[0, 1]$ is more suitable than absolute depth for PSF rendering, and they supervise the reconstruction network. Given the lack of paired metric depth for standard hyperspectral datasets, the mostly indoor scene structure, and the normalized-depth requirement of DP simulation, relative DepthAnything depth is practical for both simulation and training.

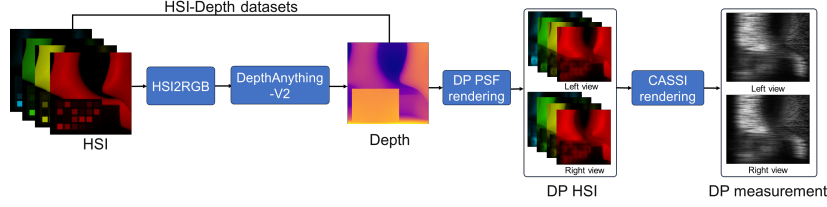


Fig. 4. Differentiable simulation pipeline for generating paired DP-CASSI measurements and depth supervision.

This pipeline is useful for two reasons. First, it lets us train the entire system with paired HSI-depth supervision while preserving the physics of both DP and CASSI formation. Second, because every operation is differentiable, it naturally supports end-to-end optimization of both the coded mask and the reconstruction network.

3.4 Joint Spectral–Depth Reconstruction Network

Given the dual-pixel measurement pair $(\mathbf{Y}^L, \mathbf{Y}^R)$, our goal is to directly reconstruct the hyperspectral image cube $\hat{\mathbf{X}} \in \mathbb{R}^{H \times W \times L}$ and the depth map $\hat{\mathbf{D}} \in \mathbb{R}^{H \times W}$ in a single forward pass. To this end, we adopt a lightweight encoder–decoder architecture SpecNet, inspired by SPECAT, while modifying it for joint spectral–depth recovery from DP-CASSI measurements.

Instead of using separate task-specific branches, our network performs *shared reconstruction* for both modalities. The input measurements are first projected into a latent feature space through a convolutional embedding layer. In parallel, a second embedding layer produces a guidance feature, which plays the role of an auxiliary mask for modulating subsequent attention blocks. The main feature stream and the guidance stream are then processed jointly throughout the network.

The encoder consists of a sequence of cumulative attention blocks (CABs) [26] followed by feature downsampling. At each stage, the feature map is updated by a CAB conditioned on the current guidance feature, after which both the feature map and the guidance feature are downsampled and propagated to the next scale. In this way, the network maintains a hierarchy of feature representations together with scale-consistent guidance signals. The bottleneck is implemented by another CAB operating at the coarsest resolution, allowing the model to aggregate global contextual information before reconstruction. The decoder progressively upsamples the latent representation back to the original spatial resolution. At each decoding stage, the upsampled feature is concatenated with the corresponding encoder feature through a skip connection and fused by a 1×1 convolution, after which a CAB further refines the fused representation under the guidance feature from the same scale. This design preserves high-resolution spatial details while enabling long-range spectral–spatial interactions across scales. Finally, a 3×3 convolutional mapping layer projects the decoded feature into $L + 1$ output channels. The first channel is interpreted as the reconstructed depth map $\hat{\mathbf{D}}$, while the remaining L channels correspond to the reconstructed hyperspectral cube $\hat{\mathbf{X}}$. In our implementation,

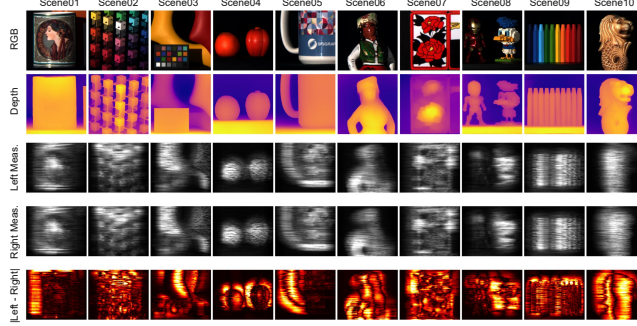


Fig. 5. Examples from the simulated DP-HSI dataset: RGB, depth, left/right measurements, and their difference.

$L = 28$. Therefore, unlike conventional multi-task architectures that split into separate decoders, our model uses a *shared encoder-decoder trunk with joint output projection*, which encourages the network to learn a unified latent representation for spectral dispersion removal and disparity-aware depth inference. This unified design is particularly suitable for DP-CASSI because the two reconstruction targets are strongly coupled in the measurements. Depth cues help explain disparity and defocus patterns, while spectral structure helps disambiguate coded dispersion. By reconstructing both quantities jointly from shared features, the network can better exploit the mutual constraints between geometry and hyperspectral appearance.

We optimize the entire system end-to-end using a weighted sum of spectral reconstruction and depth estimation losses:

$$\mathcal{L} = \lambda_{\text{hsi}} \mathcal{L}_{\text{hsi}}(\hat{\mathbf{X}}, \mathbf{X}) + \lambda_{\text{depth}} \mathcal{L}_{\text{depth}}(\hat{\mathbf{D}}, \mathbf{D}). \quad (10)$$

We use a pixel-wise spectral loss for HSI supervision and an ℓ_1 depth loss for geometry. The weights λ_{hsi} and λ_{depth} balance the two supervision terms.

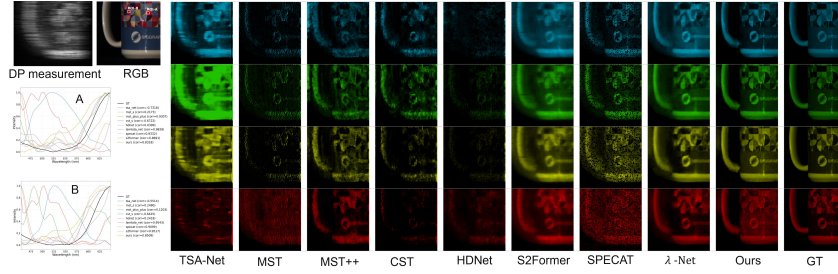
4 Experiments

4.1 Datasets, Baselines, and Metrics

Datasets and simulation setup. Examples of the simulated data are shown in Fig. 5. We evaluate the proposed method on simulated spectral-depth scenes constructed from hyperspectral images and dense geometry. Following common CASSI evaluation practice, we use the CAVE hyperspectral dataset for training and select 10 scenes from the KAIST dataset for testing. The CAVE dataset contains 32 hyperspectral images with spatial resolution 512×512 , while the KAIST dataset provides outdoor scenes with 28 spectral bands. For testing, all HSI cubes are cropped or resized to 256×256 to match the DP-CASSI simulation and reconstruction setting. This train/test protocol follows the same CAVE-to-KAIST split commonly adopted in prior CASSI works [6].

Table 1. HSI reconstruction on 10 KAIST scenes. Each cell reports PSNR/SSIM.

Method	Venue	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg
λ -Net	ICCV_2019	25.3044 0.6943	19.5710 0.5298	20.1495 0.6455	32.7445 0.8288	21.5909 0.6475	19.4997 0.5719	18.0648 0.5894	22.0526 0.5666	25.0086 0.6965	20.6381 0.4876	22.4624 0.6258
TSA-Net	ECCV_2020	15.5215 0.4699	11.7406 0.2778	9.8313 0.3152	16.6206 0.4991	11.6483 0.3730	16.2488 0.4463	10.1348 0.2609	15.5200 0.3746	12.3527 0.3881	14.5018 0.3780	13.4120 0.3783
CST-S	ECCV_2022	20.4341 0.4105	16.6243 0.3627	13.2113 0.3272	23.7816 0.6118	16.8710 0.3989	18.2945 0.5418	14.5594 0.3269	20.1619 0.5497	17.0347 0.4409	18.2025 0.4589	17.9175 0.4429
MST-S	CVPR_2022	16.1759 0.2348	13.2116 0.1471	10.3449 0.1281	19.5082 0.4357	12.1888 0.1734	16.4209 0.3542	10.2272 0.1165	16.2658 0.2836	13.8855 0.2280	14.6221 0.2975	14.2851 0.2399
MST++	CVPRW_2022	21.7670 0.5090	16.0950 0.3805	13.2079 0.3739	22.2264 0.6216	17.2104 0.4071	19.0967 0.5967	14.6797 0.3552	21.0413 0.5562	17.6213 0.4932	19.2226 0.4566	18.2168 0.4750
HDNet	CVPR_2022	16.5969 0.3145	13.0681 0.2102	9.9644 0.2107	20.9185 0.5778	10.6176 0.2217	16.9301 0.4358	9.4632 0.2283	17.3494 0.4139	13.9084 0.3045	14.2973 0.2097	14.3114 0.3127
SPECAT	CVPR_2024	21.4311 0.4815	17.0286 0.3125	14.6433 0.2615	25.2672 0.6206	17.4134 0.3139	18.8319 0.5186	14.7990 0.2661	20.4033 0.4793	17.6952 0.4199	19.0238 0.4341	18.6537 0.4108
S2Former	TPAMI_2025	25.1052 0.6875	19.2680 0.5314	18.2529 0.5963	27.9199 0.7756	21.4720 0.6361	19.8482 0.5829	17.2132 0.5713	22.2925 0.5427	22.6612 0.6750	21.1969 0.4833	21.5230 0.6082
Ours	—	26.7638 0.8090	20.0498 0.5821	17.2664 0.6583	29.2363 0.8118	24.2093 0.7841	22.8698 0.6434	19.8872 0.6647	24.2259 0.7173	19.7386 0.7017	22.8289 0.5912	22.7076 0.6964

**Fig. 6.** HSI reconstruction comparison on Scene05, including four spectral bands and representative spectral curves.

To generate realistic DP-CASSI measurements, we use the differentiable pipeline shown in Fig. 4. Specifically, the latent HSI cube is first rendered into left and right dual-pixel hyperspectral views using the DP-PSF model described in Sec. 3.2. In our implementation, the DP renderer uses 2000 discretized depth layers, a maximum depth threshold of 250, and a maximum circle-of-confusion radius of 40. The Butterworth kernel parameters are fixed to filter order 3, cutoff factor 2.5, $\beta = 0.1$, and smoothing strength 7, with an empirical CoC scale of 40.8. For each sample, the physical focus distance is determined from the median normalized depth and used to render the left and right DP hyperspectral views. These two DP views are then passed through the CASSI forward model to generate the final coded measurement pairs.

Implementation details and baselines. The network is implemented in PyTorch and trained for 300 epochs on a single NVIDIA RTX 4090 GPU. We use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, together with a cosine annealing learning-rate schedule. Unless otherwise stated, the network input is formed by stacking the left and right coded DP measurements after shift-back alignment, and the network output consists of

Table 2. Depth RMSE on 10 KAIST scenes. Lower is better.

Method	Variant	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg
λ -Net	+StereoSGBM	0.3120	0.2125	0.2526	0.3394	0.2500	0.3458	0.2036	0.2968	0.2937	0.3459	0.2852
	+DP-defocus	0.3119	0.2133	0.2508	0.3474	0.2482	0.3454	0.2015	0.2971	0.2940	0.3448	0.2854
SPECAT	+StereoSGBM	0.3120	0.2138	0.2529	0.3474	0.2505	0.3458	0.2037	0.2970	0.2943	0.3462	0.2864
	+DP-defocus	0.3093	0.2137	0.2472	0.3466	0.2466	0.3119	0.2037	0.2970	0.2923	0.3228	0.2791
S2Former	+StereoSGBM	0.3120	0.2138	0.2529	0.3474	0.2505	0.3459	0.2037	0.2971	0.2943	0.3463	0.2864
	+DP-defocus	0.3058	0.2138	0.2528	0.3439	0.2479	0.3197	0.2037	0.2957	0.2904	0.3294	0.2803
Ours	End-to-end	0.1419	0.1541	0.1103	0.0992	0.1702	0.1090	0.0794	0.0955	0.1576	0.0860	0.1203

one depth channel and 28 spectral channels. The coded aperture is parameterized as a learnable amplitude mask and optimized jointly with the reconstruction network, so that the sensing pattern is directly adapted to the joint spectral–depth recovery objective. The entire system, including the coded mask and the reconstruction backbone, is trained end-to-end.

We compare our method against representative end-to-end CASSI reconstruction networks, including λ -Net [17], TSA-Net [16], MST [4], MST++ [5], CST [3], HD-Net [12], SPECAT [26], and S2Former [20]. Since these methods are originally designed for hyperspectral reconstruction only, we use them as spectral baselines under the same simulated scenes and compare their HSI predictions against our joint spectral–depth formulation.

For depth estimation, we further construct two additional two-stage baselines based on the reconstructed left/right views. Specifically, each spectral reconstruction baseline first predicts the dual-pixel hyperspectral views, which are then converted to RGB images through HSI2RGB projection. The resulting RGB pair is subsequently fed into either StereoSGBM, an OpenCV implementation of semi-global stereo matching [11], or a dual-pixel defocus-based depth estimator, denoted as *DP-defocus*, based on defocus map estimation [22]. In this way, both StereoSGBM and DP-defocus serve as post-hoc depth estimation baselines built on top of the reconstructed hyperspectral outputs, rather than directly operating on the raw DP-CASSI measurements.

Our method differs from these baselines in two important aspects. First, it explicitly models dual-pixel geometry in the forward process instead of treating depth estimation as a separate post-processing stage. Second, it jointly optimizes the coded mask and the reconstruction network for simultaneous hyperspectral and depth recovery, whereas the two-stage baselines decouple spectral reconstruction from geometric inference.

Evaluation metrics. For hyperspectral reconstruction, we report peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). For depth estimation, we report the root mean squared error (RMSE) between the predicted depth map and the ground-truth depth map. In addition, we provide qualitative depth visualizations and representative spectral curves to evaluate whether the recovered geometry is consistent with the scene structure and whether the reconstructed spectrum preserves relative amplitude and shape. Our evaluation therefore emphasizes not only spectral fidelity, but also the ability of DP-CASSI to recover meaningful geometry from the same coded snapshot.

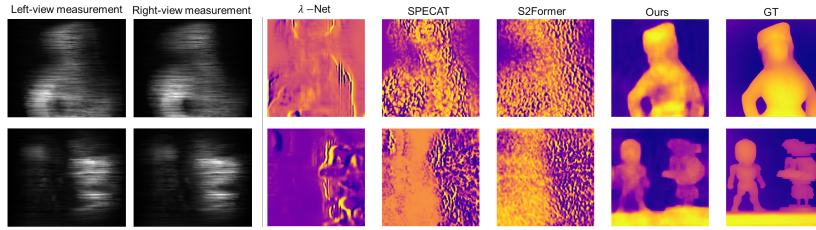


Fig. 7. Depth comparison on scenes 06 and 08. Our end-to-end prediction is structurally closer to ground truth than two-stage DP-defocus baselines.

4.2 Quantitative Comparison

Table 1 reports quantitative results on the KAIST test set. Our method achieves the best mean PSNR/SSIM and the strongest results on most scenes. The explicit use of DP geometry leads to better recovery than single-task HSI networks. Notably, λ -Net remains competitive on several PSNR entries, suggesting that plain CASSI inversion can still be effective when geometry is weakly expressed. We also observe our method does not produce the highest spectral fidelity over the baseline methods, this may be compromised by the joint learning mechanism of our E2E network. Overall, however, the table supports our main claim: explicitly modeling dual-pixel disparity inside the CASSI formation improves robustness under depth variation.

To evaluate the depth estimation quality of our system, we compare against three representative HSI reconstruction baselines: λ -Net, SPECAT, and S2Former. For each baseline, we first reconstruct the HSI from the DP-CASSI measurements, synthesize a pair of RGB images (left and right view) from the reconstructed HSI, and then feed the stereo RGB pair into a state-of-the-art dual-pixel defocus estimation method [22] to obtain a depth map. As shown in Table 2, our method achieves a mean RMSE of 0.1203 in 10 test scenes, outperforming λ -Net (0.2854), SPECAT (0.2791) and S2Former (0.2803) by a large margin – reducing the depth RMSE by more than 57% compared to the best baseline. This improvement demonstrates that our joint DP-CASSI reconstruction framework, which simultaneously recovers spectral and depth information from the raw dual-pixel measurement, substantially outperforms the cascade approach of independent HSI reconstruction followed by off-the-shelf depth estimation.

4.3 Qualitative Results

Figure 6 compares reconstructed spectral bands on Scene 5 of KAIST test dataset. The baselines frequently show either spectral bleeding or weak local contrast, especially around object boundaries where depth variation is more pronounced. In contrast, our method better preserves the shape of fine structures and yields more stable spectral profiles. The spectral curves shown at the bottom-left further confirm that the proposed DP-CASSI model preserves both relative spatial and spectral trend more faithfully than baseline methods.

Fig. 7 visualizes the estimated depth maps for scene 06 and scene 08. The three baseline methods (λ -Net, SPECAT, S2Former) produce highly noisy and structurally

Table 3. Mean HSI quality on left-view (DP_L), right-view (DP_P) and ours.

Metric	DP_L	DP_R	Ours
PSNR $\uparrow$	24.0457	23.4629	22.7076
SSIM $\uparrow$	0.6992	0.6826	0.6964

**Fig. 8.** RGB comparison synthesized from DP spectrum, our reconstruction, and GT.

incoherent depth maps, as the reconstructed spectrum lacks the fine spatial fidelity required for accurate defocus cue extraction. In contrast, our method yields a clean, well-structured depth map that clearly delineates the object silhouette and spatial layout, closely matching the ground-truth depth. This qualitative comparison further confirms that relying on HSI reconstruction as an intermediate step introduces compounding errors that degrade depth estimation quality, whereas our end-to-end approach directly exploits the dual-pixel disparity encoded in the DP-CASSI measurement.

4.4 Ablation Study

Table 3 and Fig. 8 demonstrate that the dual-pixel input spectrum is already blurred relative to the ground truth, making our task harder than standard flat-scene CASSI reconstruction. The first two RGB images are synthesized from simulated left/right CASSI hyperspectral views and already contain spatially varying DP blur. The third image is our reconstruction, which shows visibly reduced blur while recovering spectrum.

Table 4 evaluates dataset generalization beyond KAIST. On hyperspectral datasets CAVE and ICVL, our method consistently improves PSNR, SSIM, SAM, and RMSE over MST. SAM (spectral angle mapper) measures the angle between reconstructed and ground-truth spectral vectors, with smaller angles indicating higher spectral similarity.

Table 5 compares the proposed learned coded aperture against a fixed mask. The learned mask consistently improves hyperspectral reconstruction quality in terms of PSNR and SSIM, while also reducing the depth RMSE. This result indicates that jointly optimizing the sensing pattern together with the reconstruction network leads to a more task-adaptive optical encoding, benefiting both spectral fidelity and geometric recovery.

We also compare SpecNet with unfolding-based baselines in Table 6. DAUHST [6] and SSR [29] achieve higher PSNR and SSIM, but require much larger computation; SpecNet provides a lightweight single-pass alternative while preserving competitive structural quality.

Table 4. Additional validation on CAVE and ICVL datasets.

Dataset	Method	PSNR $\uparrow$	SSIM $\uparrow$	SAM $\downarrow$	RMSE $\downarrow$
CAVE	MST	13.84	0.17	0.87	0.26
CAVE	Ours	22.48	0.57	0.33	0.21
ICVL	MST	15.26	0.16	1.08	0.40
ICVL	Ours	22.08	0.50	0.41	0.22

Table 5. Fixed versus learned coded aperture.

Mask setting	PSNR $\uparrow$	SSIM $\uparrow$	RMSE $\downarrow$
Fixed mask	22.45	0.70	0.13
Learned mask	22.71	0.70	0.12

Table 6. Unfolding-based versus end-to-end reconstruction.

Method	Type	PSNR $\uparrow$	SSIM $\uparrow$	Params / FLOPs
DAUHST-3stg	Unfolding	23.9638	0.6820	2.08M / 27.17G
SSR-S	Unfolding	24.0606	0.6855	1.73M / 26.37G
SpecNet (ours)	E2E	22.7076	0.6964	0.29M / 12.4G

5 Conclusion

We presented DP-CASSI, a dual-pixel extension of coded aperture snapshot spectral imaging for joint spectral–depth recovery. By incorporating dual-pixel sensing into the CASSI optical path, the proposed system encodes both spectral dispersion and depth-dependent disparity in a single snapshot. We formulated a coupled forward model, developed a differentiable simulation pipeline for supervised training, and introduced a joint reconstruction network with shared spectral–geometric features. Experiments on simulated datasets show that explicitly modeling dual-pixel geometry improves both spectral fidelity and depth estimation compared with conventional single-task baselines. We hope this work encourages future snapshot spectral imagers to exploit geometric cues rather than treat them solely as a source of degradation.

Acknowledgments. This work was supported by “Pioneer” and “Leading Goose” R&D Program of Zhejiang (grant 2024SDXHDX0006, 2024C03182). This work was supported by National Key R&D Program of China (2024YFF0505603) and the National Natural Science Foundation of China (grant number U2541205, 62271414). (Corresponding author: Xin Yuan.)

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abuolaim, A., Delbracio, M., Kelly, D., Brown, M.S., Milanfar, P.: Learning to reduce defocus blur by realistically modeling dual-pixel data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2289–2298 (2021)

2. Baek, S.H., Ikoma, H., Jeon, D.S., Li, Y., Heidrich, W., Wetzstein, G., Kim, M.H.: Single-shot hyperspectral-depth imaging with learned diffractive optics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2651–2660 (2021)
3. Cai, Y., Lin, J., Hu, X., Wang, H., Yuan, X., Zhang, Y., Timofte, R., Van Gool, L.: Coarse-to-fine sparse transformer for hyperspectral image reconstruction. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. LNCS, vol. 13677, pp. 686–704. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19790-1_41
4. Cai, Y., Lin, J., Hu, X., Wang, H., Yuan, X., Zhang, Y., Timofte, R., Van Gool, L.: Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17502–17511 (2022)
5. Cai, Y., Lin, J., Lin, Z., Wang, H., Zhang, Y., Pfister, H., Timofte, R., Van Gool, L.: MST++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 745–755 (2022)
6. Cai, Y., Lin, J., Wang, H., Yuan, X., Ding, H., Zhang, Y., Timofte, R., Van Gool, L.: Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 37749–37761. Curran Associates, Inc. (2022). <https://doi.org/10.52202/068431-2736>
7. Cui, Q., Park, J., Smith, R.T., Gao, L.: Snapshot hyperspectral light field imaging using image mapping spectrometry. *Optics Letters* **45**(3), 772–775 (2020). <https://doi.org/10.1364/OL.382088>
8. Ding, X., Hu, L., Zhou, S., Wang, X., Li, Y., Han, T., Lu, D., Che, G.: Snapshot depth-spectral imaging based on image mapping and light field. *EURASIP Journal on Advances in Signal Processing* **2023**, 24 (2023). <https://doi.org/10.1186/s13634-023-00983-7>
9. Fu, Y., Zheng, Y., Sato, I., Sato, Y.: Exploiting spectral-spatial correlation for coded hyperspectral image restoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3727–3736 (2016)
10. Garg, R., Wadhwa, N., Ansari, S., Barron, J.T.: Learning single camera depth estimation using dual-pixels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7628–7637 (2019)
11. Hirschmüller, H.: Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **30**(2), 328–341 (2008). <https://doi.org/10.1109/TPAMI.2007.1166>
12. Hu, X., Cai, Y., Lin, J., Wang, H., Yuan, X., Zhang, Y., Timofte, R., Van Gool, L.: HDNet: High-resolution dual-domain learning for spectral compressive imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17542–17551 (2022)
13. Kurita, T., Kondo, Y., Sun, L., Sasaki, T., Nitta, S., Hashimoto, Y., Muramatsu, Y., Moriuchi, Y.: Revisiting disparity from dual-pixel images: Physics-informed lightweight depth estimation. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). pp. 8378–8388 (2025)
14. Liu, Y., Yuan, X., Suo, J., Brady, D.J., Dai, Q.: Rank minimization for snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(12), 2990–3006 (2019). <https://doi.org/10.1109/TPAMI.2018.2873587>
15. Luo, J., Forsberg, E., Fu, S., Xing, Y., Liao, J., Jiang, J., Zheng, Y., He, S.: 4D dual-mode staring hyperspectral-depth imager for simultaneous spectral sensing and surface shape measurement. *Optics Express* **30**(14), 24804–24821 (2022). <https://doi.org/10.1364/OE.460412>

16. Meng, Z., Ma, J., Yuan, X.: End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) *Computer Vision – ECCV 2020*. LNCS, vol. 12368, pp. 187–204. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58592-1_12
17. Miao, X., Yuan, X., Pu, Y., Athitsos, V.: λ -Net: Reconstruct hyperspectral images from a snapshot measurement. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 4059–4069 (2019)
18. Pan, L., Chowdhury, S., Hartley, R., Liu, M., Zhang, H., Li, H.: Dual pixel exploration: Simultaneous depth estimation and image restoration. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 4340–4349 (2021)
19. Shi, Z., Chugunov, I., Bijelic, M., Côté, G., Yeom, J., Fu, Q., Amata, H., Heidrich, W., Heide, F.: Split-aperture 2-in-1 computational cameras. *ACM Transactions on Graphics* **43**(4), 141:1–141:19 (2024). <https://doi.org/10.1145/3658225>
20. Wang, J., Li, K., Zhang, Y., Yuan, X., Tao, Z.: S^2 -transformer for mask-aware hyperspectral image reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **47**(6), 4299–4316 (2025). <https://doi.org/10.1109/TPAMI.2025.3543842>
21. Wang, L., Xiong, Z., Shi, G., Zeng, W., Wu, F.: Simultaneous depth and spectral imaging with a cross-modal stereo system. *IEEE Transactions on Circuits and Systems for Video Technology* **28**(3), 812–817 (2018)
22. Xin, S., Wadhwa, N., Xue, T., Barron, J.T., Srinivasan, P.P., Chen, J., Gkioulekas, I., Garg, R.: Defocus map estimation and deblurring from a single dual-pixel image. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 2228–2238 (2021)
23. Xiong, Z., Wang, L., Li, H., Liu, D., Wu, F.: Snapshot hyperspectral light field imaging. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 3270–3278 (2017)
24. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything V2. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) *Advances in Neural Information Processing Systems*. vol. 37, pp. 21875–21911. Curran Associates, Inc. (2024). <https://doi.org/10.52202/079017-0688>
25. Yao, M., Xiong, Z., Wang, L., Liu, D., Chen, X.: Spectral-depth imaging with deep learning based reconstruction. *Optics Express* **27**(26), 38312–38325 (2019). <https://doi.org/10.1364/OE.27.038312>
26. Yao, Z., Liu, S., Yuan, X., Fang, L.: SPECAT: SPatial-spEctral cumulative-attention transformer for high-resolution hyperspectral image reconstruction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 25368–25377 (2024)
27. Yuan, X.: Generalized alternating projection based total variation minimization for compressive sensing. In: *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. pp. 2539–2543 (2016). <https://doi.org/10.1109/ICIP.2016.7532817>
28. Yuan, X., Brady, D.J., Katsaggelos, A.K.: Snapshot compressive imaging: Theory, algorithms, and applications. *IEEE Signal Processing Magazine* **38**(2), 65–88 (2021). <https://doi.org/10.1109/MSP.2020.3023869>
29. Zhang, J., Zeng, H., Chen, Y., Yu, D., Zhao, Y.P.: Improving spectral snapshot reconstruction with spectral-spatial rectification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 25817–25826 (2024)